Health Care project

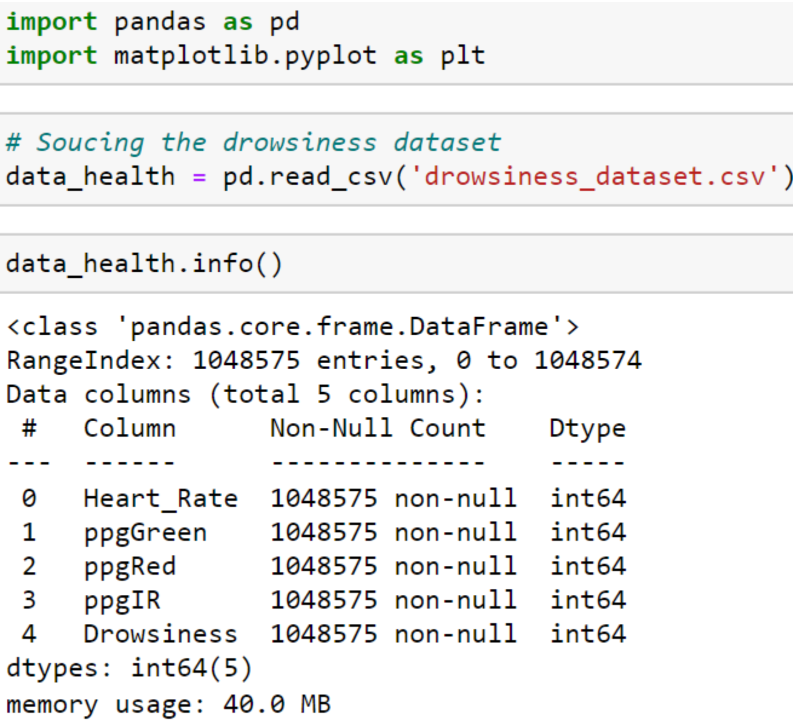
EXPLORATORY DATA ANALYSIS ON THE DROWSINESS LEVELS.

Introduction

- **Scenario**: Wearable technology company that produces smart watches emits different signals, which include variations in green, red and infrared light. One of the key features is that its ability to detect and alert users to potential drowsiness based on the physiological data.

- **Objective** : To perform **an Exploratory Data Analysis (EDA)** on a dataset collected from these smartwatches. The dataset includes various physiological parameters along with a 'drowsiness' label, which indicates the level of sleepiness based on an adapted Karolinska Sleepiness Scale (KSS).

Knowing about the data



```
import pandas as pd
import matplotlib.pyplot as plt
```

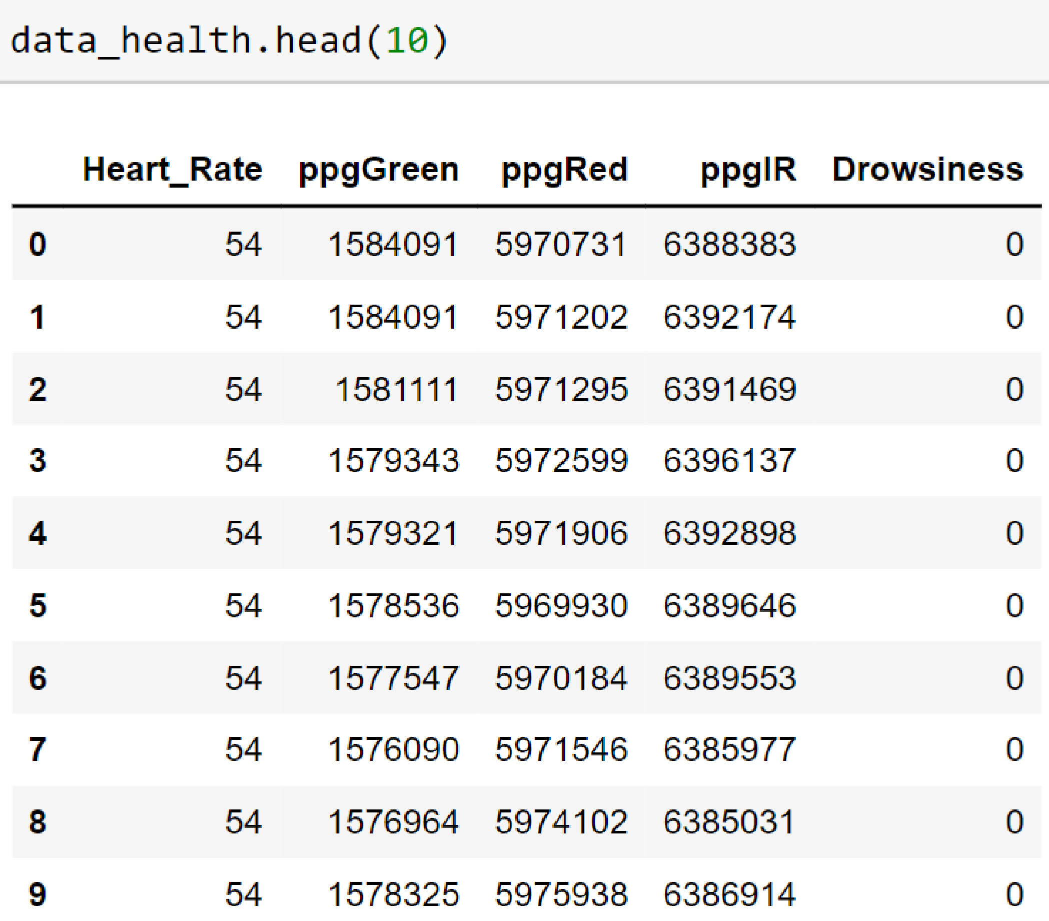
```
# Soucing the drowsiness dataset
data_health = pd.read_csv('drowsiness_dataset.csv')
```

```
data_health.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 5 columns):
 #   Column      Non-Null Count    Dtype
---  ------      --------------    -----
 0   Heart_Rate  1048575 non-null  int64
 1   ppgGreen    1048575 non-null  int64
 2   ppgRed      1048575 non-null  int64
 3   ppgIR       1048575 non-null  int64
 4   Drowsiness  1048575 non-null  int64
dtypes: int64(5)
memory usage: 40.0 MB
```

- The dataset has got 5 columns with no missing values.
- All the 5 columns are numerical type with integer values.
- The dataset has got 1,048,575 rows (around 1.5 million rows)

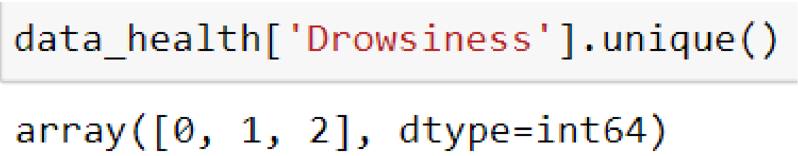
Source Dataset



```
data_health.head(10)
```

	Heart_Rate	ppgGreen	ppgRed	ppgIR	Drowsiness
0	54	1584091	5970731	6388383	0
1	54	1584091	5971202	6392174	0
2	54	1581111	5971295	6391469	0
3	54	1579343	5972599	6396137	0
4	54	1579321	5971906	6392898	0
5	54	1578536	5969930	6389646	0
6	54	1577547	5970184	6389553	0
7	54	1576090	5971546	6385977	0
8	54	1576964	5974102	6385031	0
9	54	1578325	5975938	6386914	0

First 10 rows of the Dataset



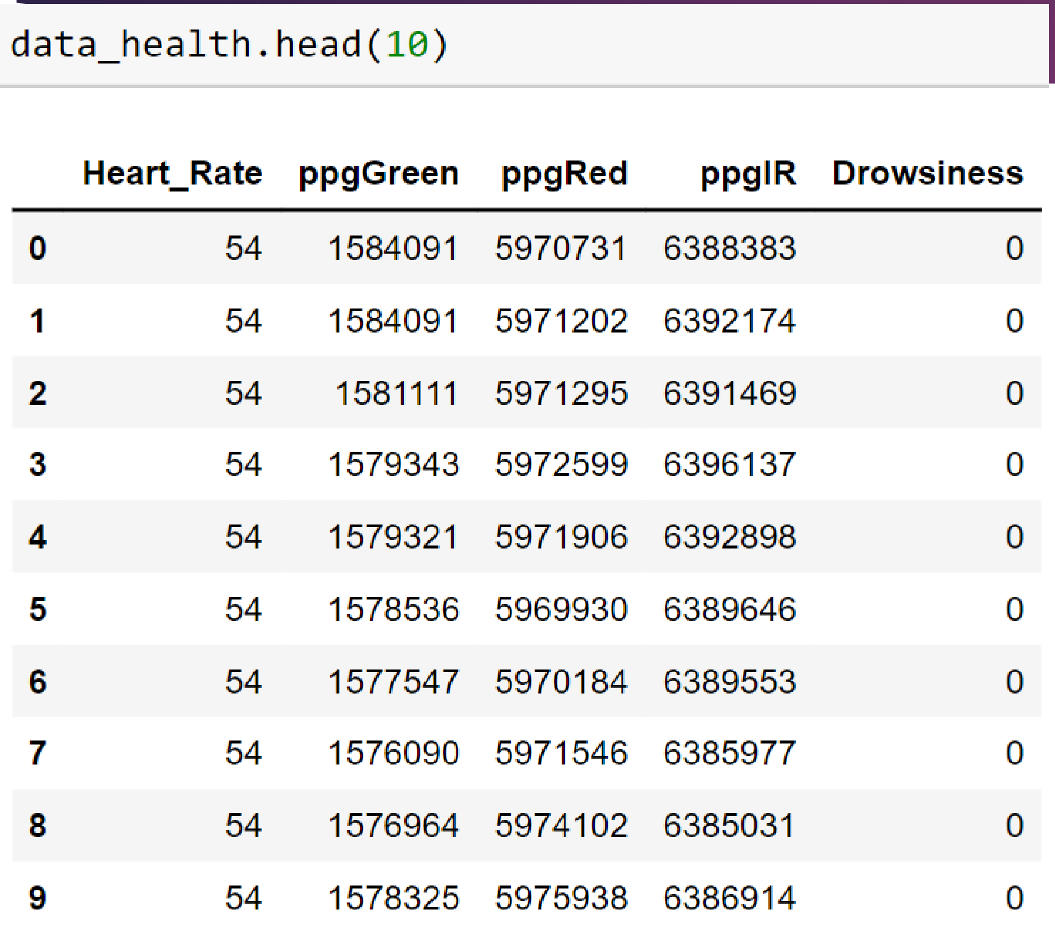
```
data_health['Drowsiness'].unique()
```

```
array([0, 1, 2], dtype=int64)
```

The Drowsiness column has only 3 unique values:

0 represents alertness.
1 moderately alert.
2 represents significant drowsiness.(extremely sleepy)

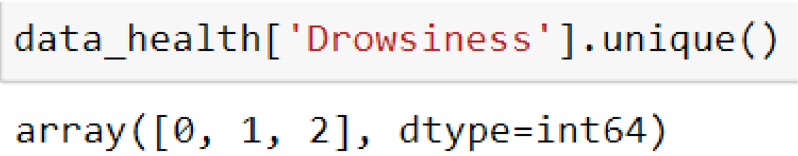
Source Dataset



```
data_health.head(10)
```

	Heart_Rate	ppgGreen	ppgRed	ppgIR	Drowsiness
0	54	1584091	5970731	6388383	0
1	54	1584091	5971202	6392174	0
2	54	1581111	5971295	6391469	0
3	54	1579343	5972599	6396137	0
4	54	1579321	5971906	6392898	0
5	54	1578536	5969930	6389646	0
6	54	1577547	5970184	6389553	0
7	54	1576090	5971546	6385977	0
8	54	1576964	5974102	6385031	0
9	54	1578325	5975938	6386914	0

First 10 rows of the Dataset



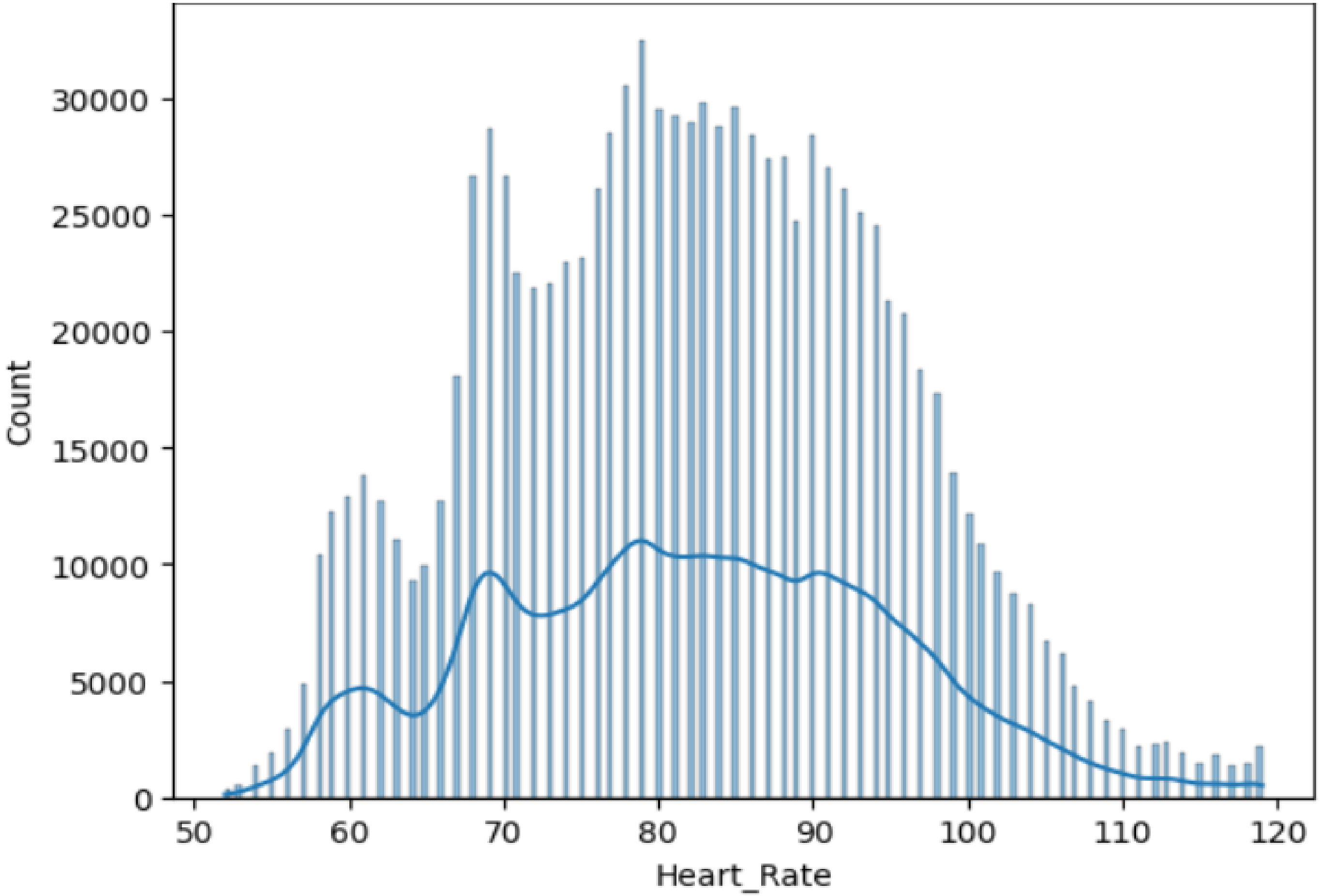
```
data_health['Drowsiness'].unique()
```

```
array([0, 1, 2], dtype=int64)
```

The Drowsiness column has only 3 unique values:

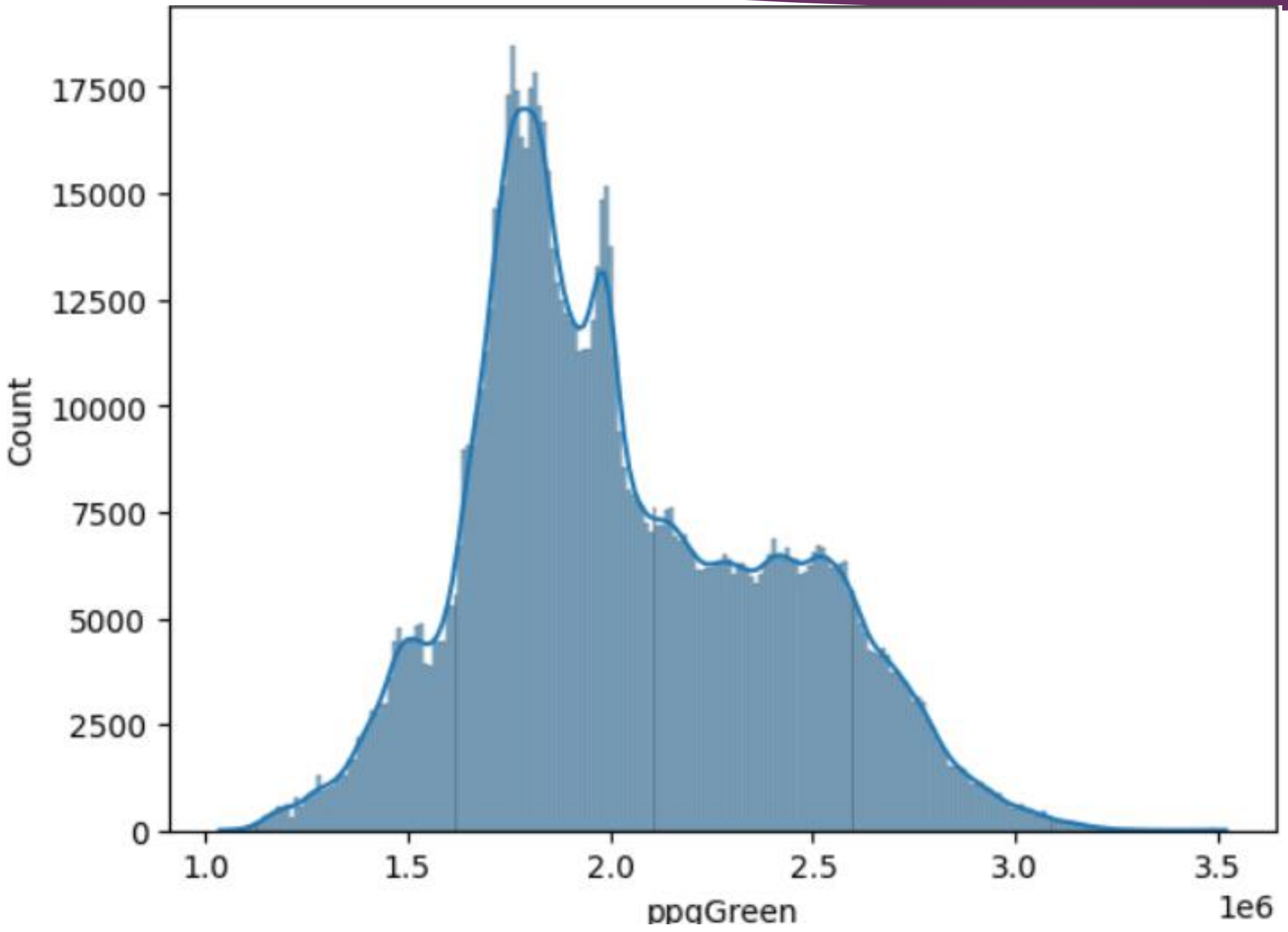
0 represents alertness.
1 moderately alert.
2 represents significant drowsiness.(extremely sleepy)

Histograms – Heart rate Column



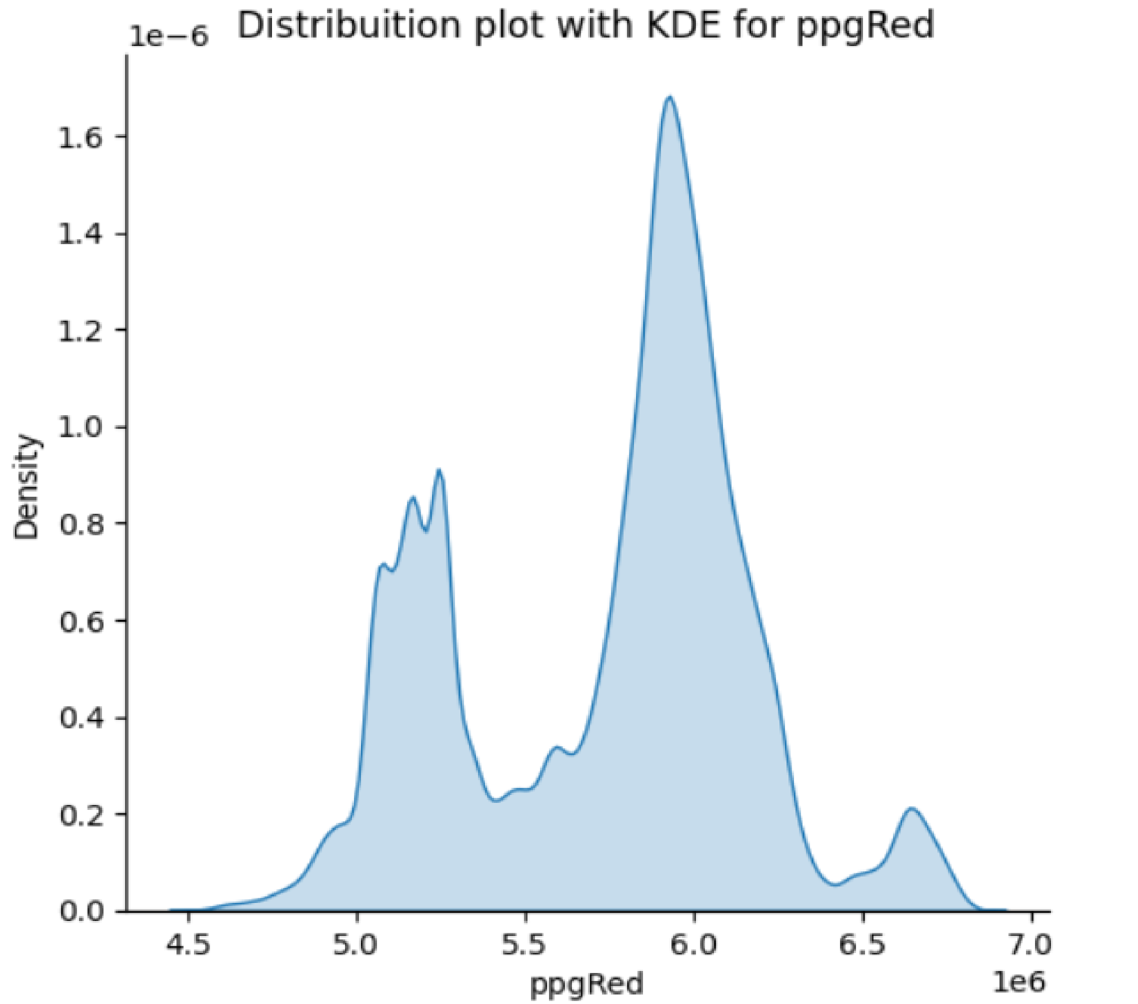
- The Heart rate values are normally distributed.
- Majority of the values are around the mean value of 82

Histograms -ppgGreen



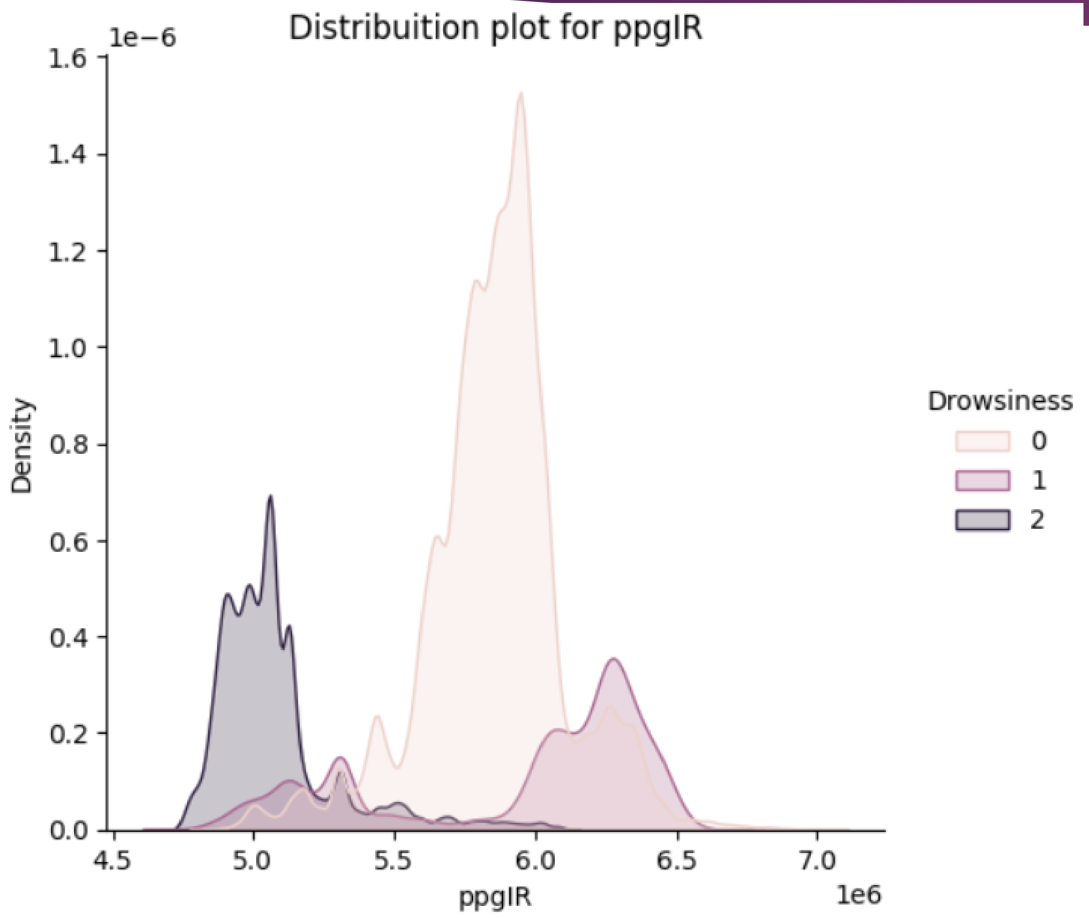
- The ppgGreen values are normally distributed.
- Majority of the values are in between 1.5M and 2M

Histograms -ppgRed



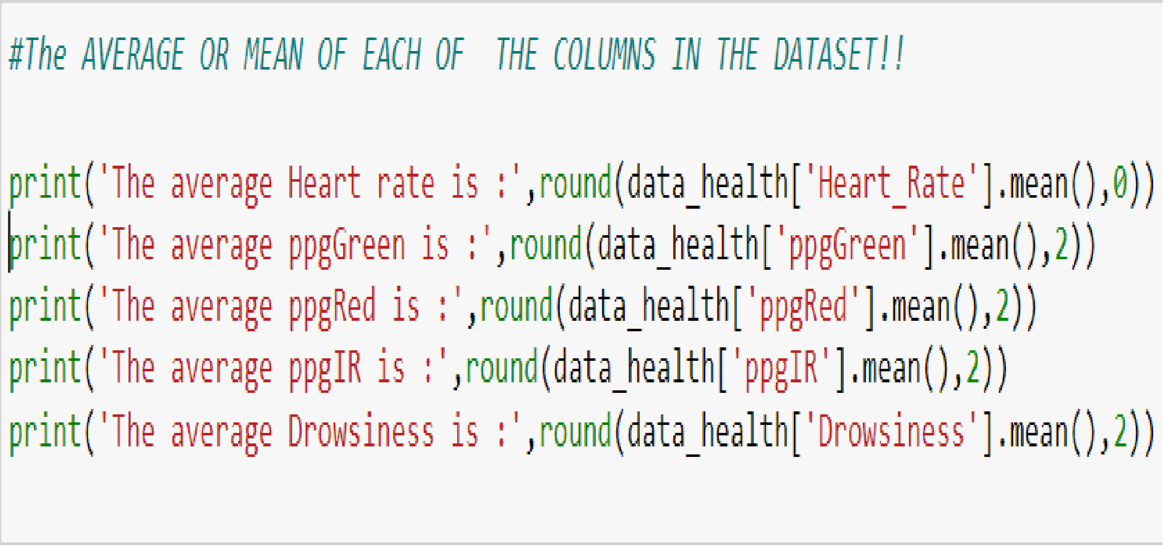
- The ppgRed values are normally distributed.
- Majority of the values are in between 5.75 M and 6.25 M

Distribution Plot –KDE plot

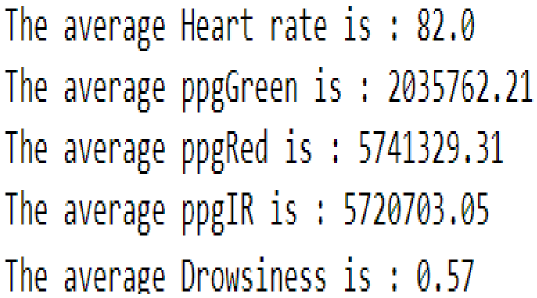


- For alert individuals, the ppgIR values are normally distributed with its values between 5.5M and 6.25M
- For drowsy individuals, their ppgIR values vary between 6M and 6.5M
- For significantly drowsy individuals the ppgIR values are between 4.75M and 5.25M,

Descriptive Statistics -Mean



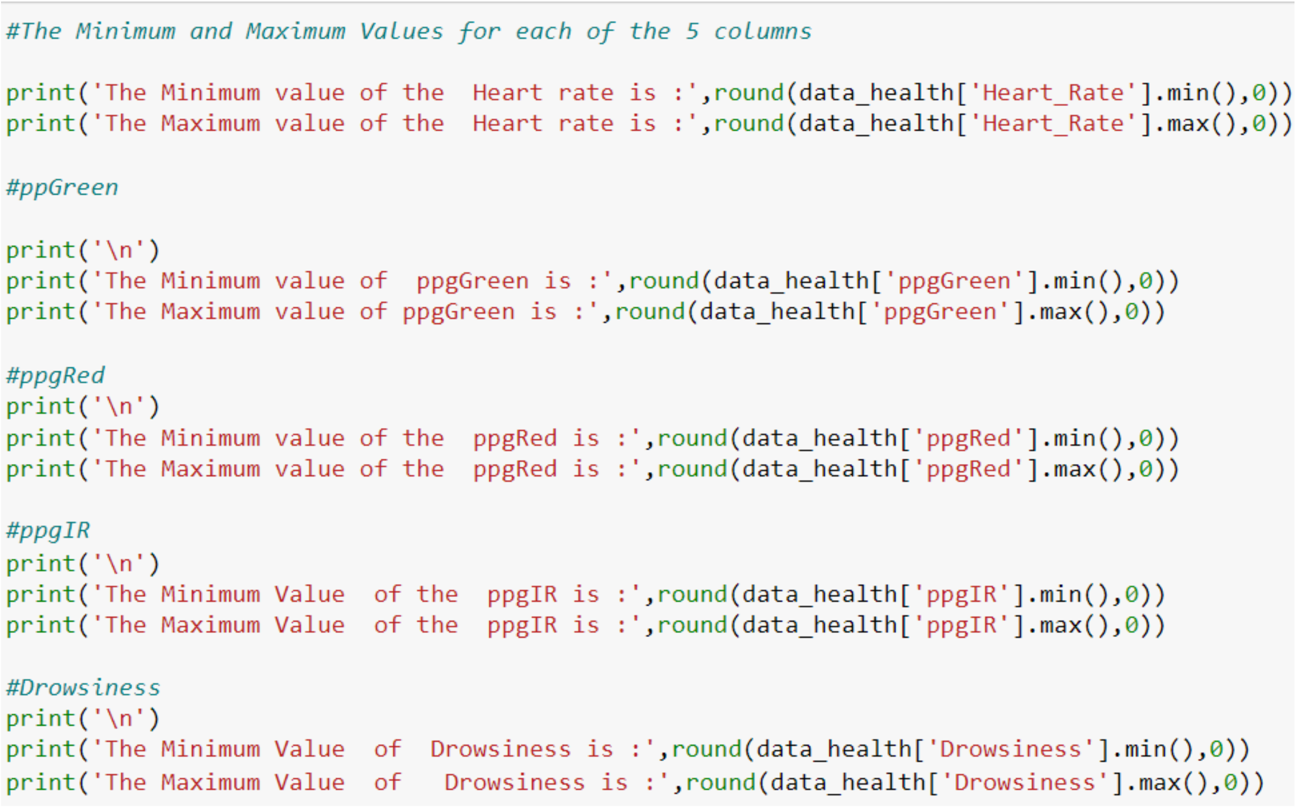
```python
#The AVERAGE OR MEAN OF EACH OF  THE COLUMNS IN THE DATASET!!

print('The average Heart rate is :',round(data_health['Heart_Rate'].mean(),0))
print('The average ppgGreen is :',round(data_health['ppgGreen'].mean(),2))
print('The average ppgRed is :',round(data_health['ppgRed'].mean(),2))
print('The average ppgIR is :',round(data_health['ppgIR'].mean(),2))
print('The average Drowsiness is :',round(data_health['Drowsiness'].mean(),2))
```



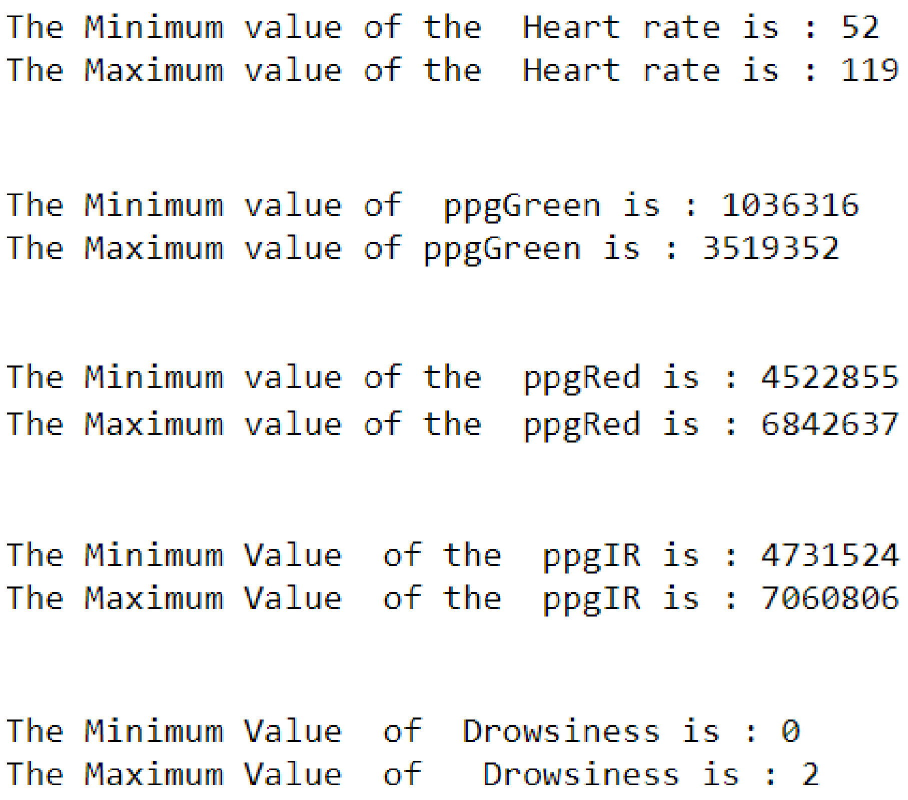
```
The average Heart rate is : 82.0
The average ppgGreen is : 2035762.21
The average ppgRed is : 5741329.31
The average ppgIR is : 5720703.05
The average Drowsiness is : 0.57
```

- The average heart rate is 82 which is normal rate.
- The average value for ppgRed and ppgIR is 5.7M, which means Red emission is dependent on the Infrared.
- The average value for ppgGreen is 2M which is much less than ppgRed and ppgIR.
- The average value for Drowsiness is 0.6 which in between 0 and 1 which means mostly the patients are alert.

Descriptive Statistics –Minimum and Maximum value

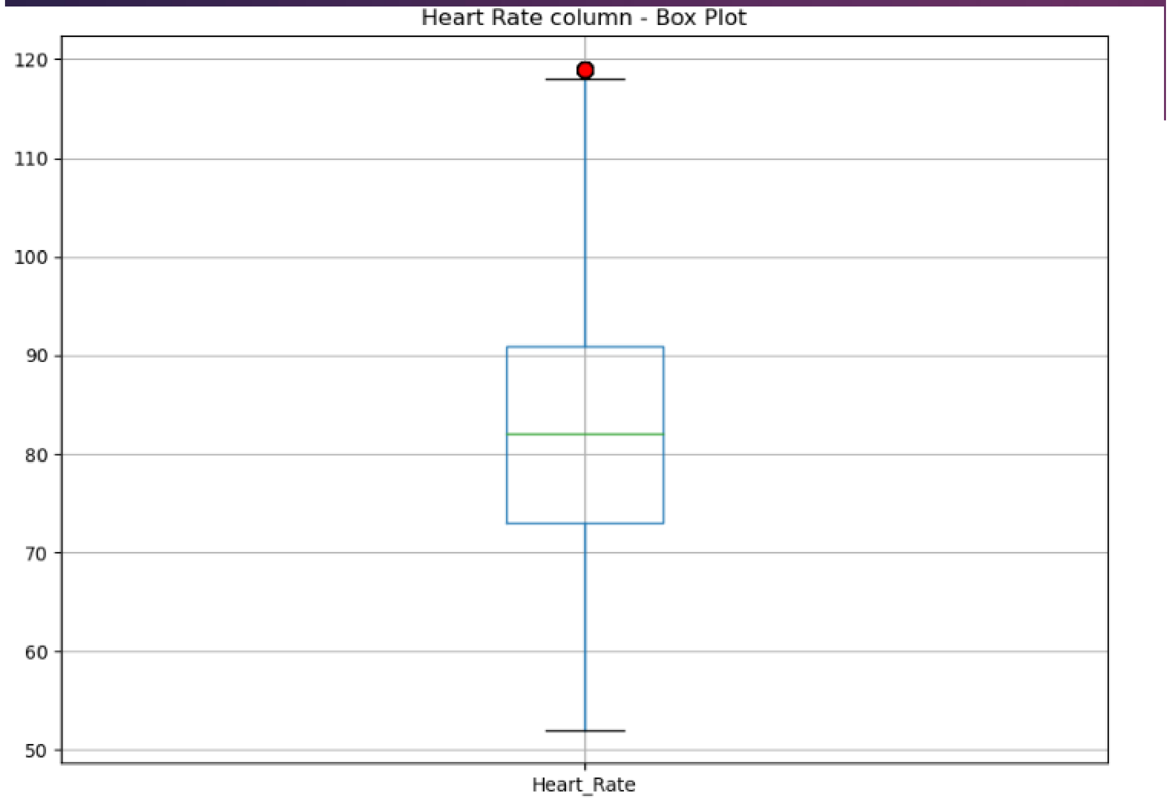


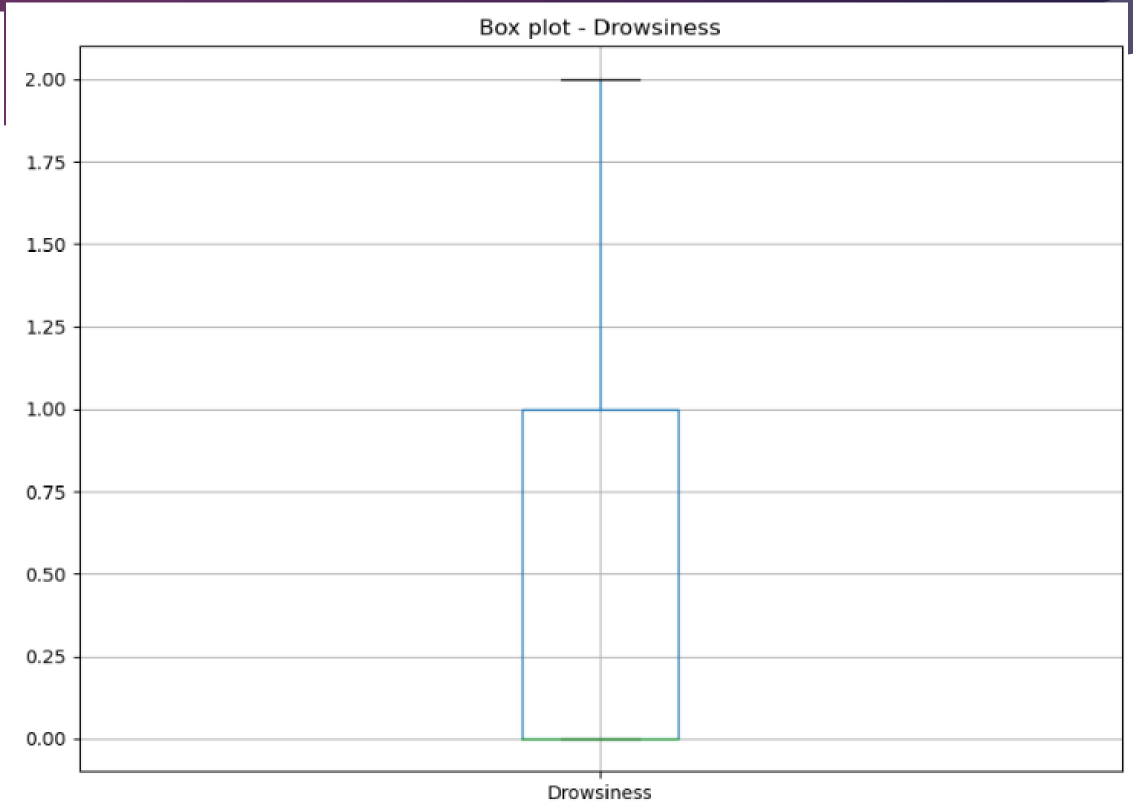
```
#The Minimum and Maximum Values for each of the 5 columns

print('The Minimum value of the  Heart rate is :',round(data_health['Heart_Rate'].min(),0))
print('The Maximum value of the  Heart rate is :',round(data_health['Heart_Rate'].max(),0))

#ppGreen

print('\n')
print('The Minimum value of  ppgGreen is :',round(data_health['ppgGreen'].min(),0))
print('The Maximum value of ppgGreen is :',round(data_health['ppgGreen'].max(),0))

#ppgRed
print('\n')
print('The Minimum value of the  ppgRed is :',round(data_health['ppgRed'].min(),0))
print('The Maximum value of the  ppgRed is :',round(data_health['ppgRed'].max(),0))

#ppgIR
print('\n')
print('The Minimum Value  of the  ppgIR is :',round(data_health['ppgIR'].min(),0))
print('The Maximum Value  of the  ppgIR is :',round(data_health['ppgIR'].max(),0))

#Drowsiness
print('\n')
print('The Minimum Value  of  Drowsiness is :',round(data_health['Drowsiness'].min(),0))
print('The Maximum Value  of   Drowsiness is :',round(data_health['Drowsiness'].max(),0))
```



```
The Minimum value of the  Heart rate is : 52
The Maximum value of the  Heart rate is : 119


The Minimum value of  ppgGreen is : 1036316
The Maximum value of ppgGreen is : 3519352


The Minimum value of the  ppgRed is : 4522855
The Maximum value of the  ppgRed is : 6842637


The Minimum Value  of the  ppgIR is : 4731524
The Maximum Value  of the  ppgIR is : 7060806


The Minimum Value  of  Drowsiness is : 0
The Maximum Value  of   Drowsiness is : 2
```

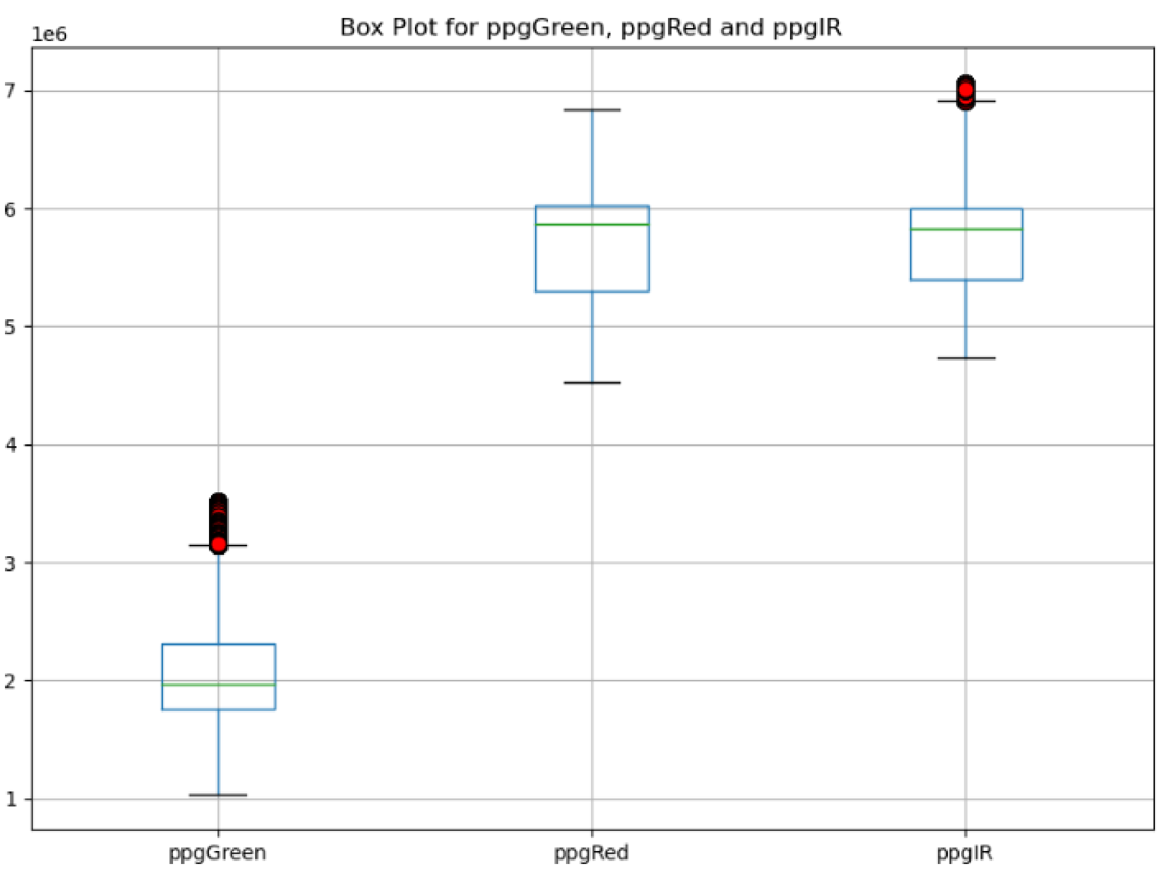
Box Plots





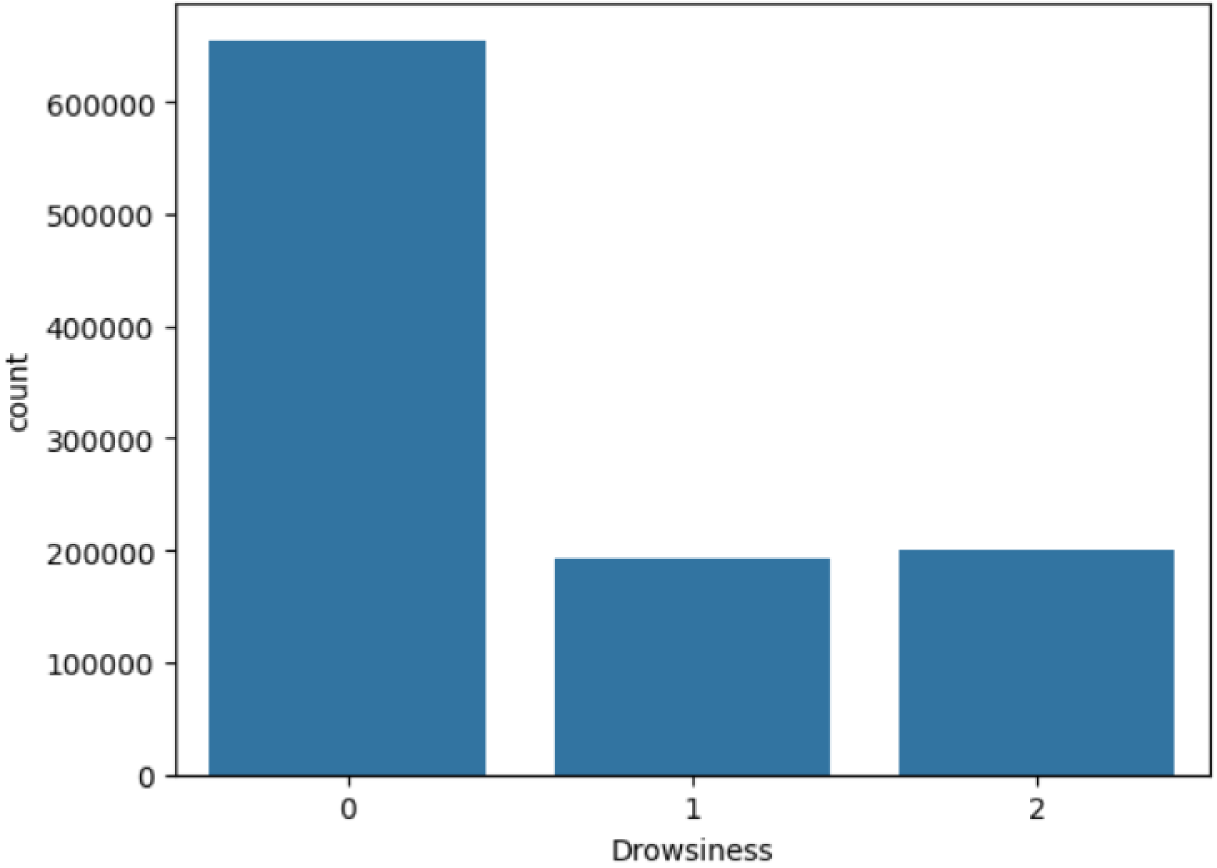
Heart rate column has outliers with the observations uniformly spread above and below the median.
The Median and the minimum value for Drowsiness is 0.
Majority of the observations are having values between 0 and 1 .

Box Plots -



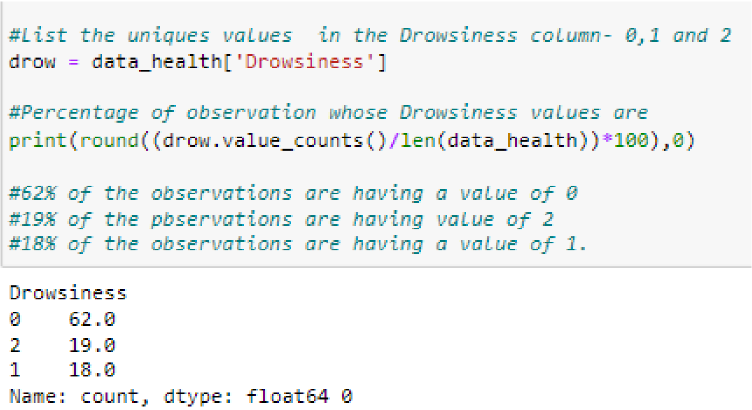
- Outliers present for ppgGreen and ppgIR.
- The box plot for ppgRed and ppgIR are similar with similar middle value as well.
- The minimum and maximum values for ppgGreen are far below when compared to ppgRed and ppgIR.
- For ppgRed and ppgIR majority of the observations are below the middle value while for ppgGreen most of the observations are above middle value for ppgGreen.
- Inter Quartile Range for ppgRed and ppgIR are similar which means they share the same range of values.

More about Drowsiness Column



- More than half of the observations(around 62%) are having value of 0 , which means most of the customers are alert.

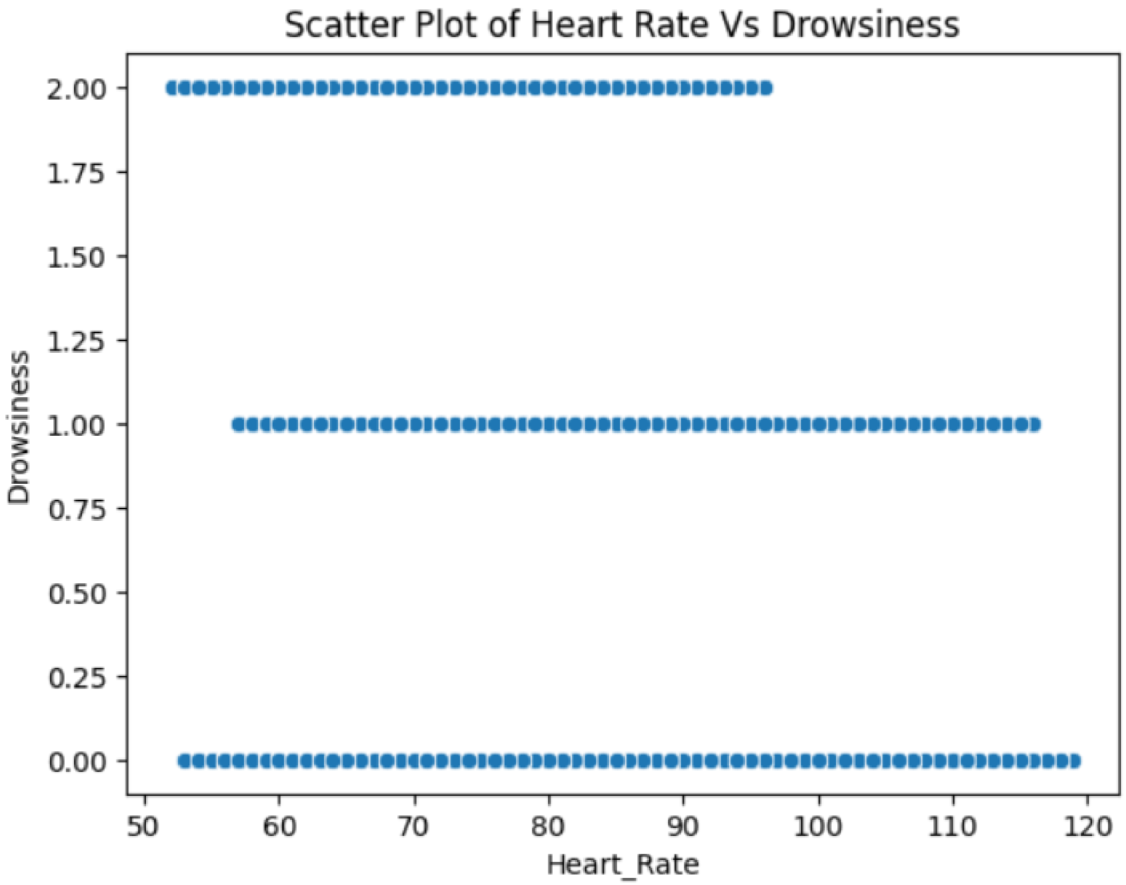
- Around 19% of them are significantly drowsy or extremely sleepy.

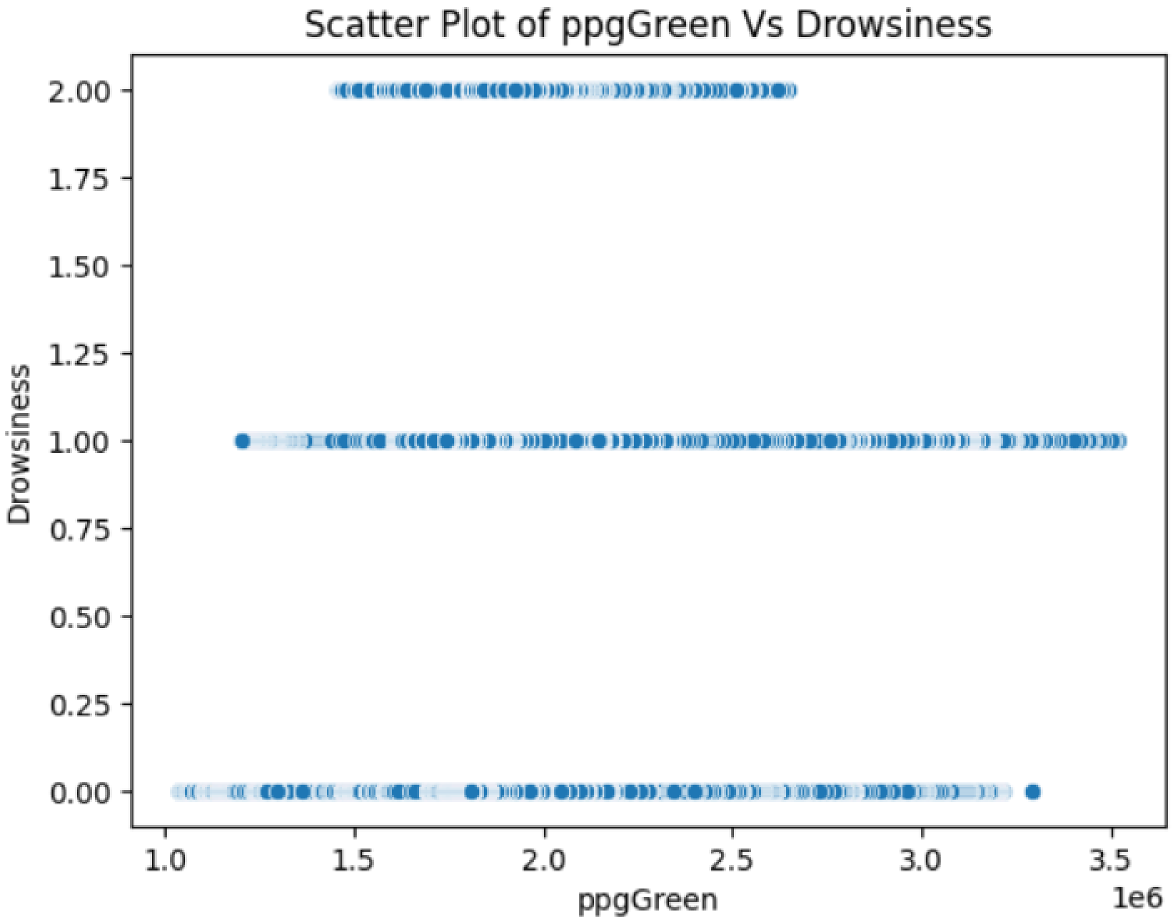


```
#List the uniques values  in the Drowsiness column- 0,1 and 2
drow = data_health['Drowsiness']

#Percentage of observation whose Drowsiness values are
print(round((drow.value_counts()/len(data_health))*100),0)

#62% of the observations are having a value of 0
#19% of the pbservations are having value of 2
#18% of the observations are having a value of 1.
```

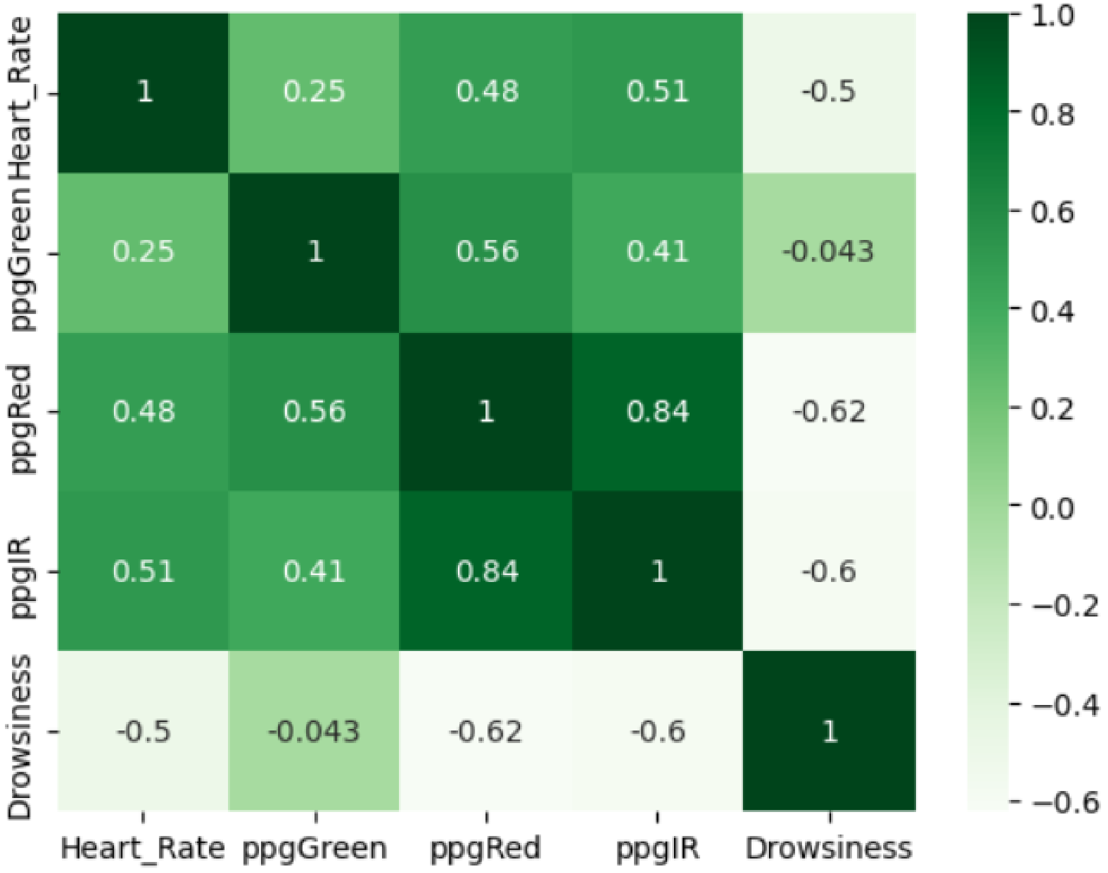
```
Drowsiness
0    62.0
2    19.0
1    18.0
Name: count, dtype: float64 0
```

Scatter Plot

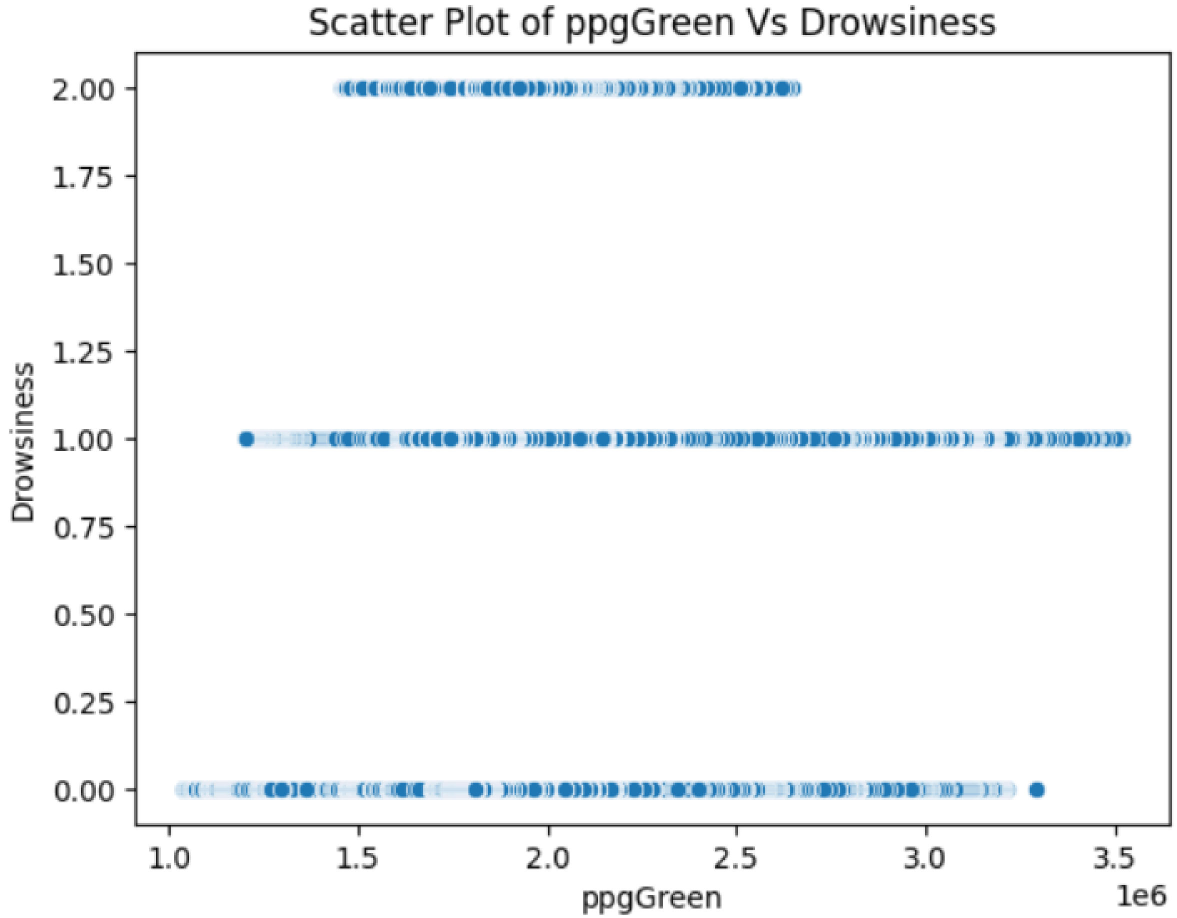




Correlation Matrix

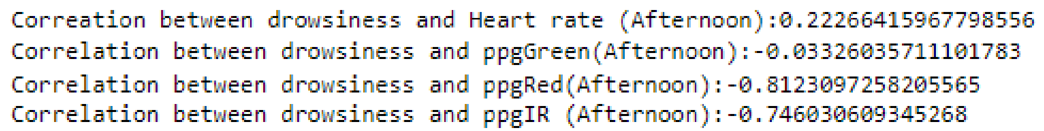


	Heart_Rate	ppgGreen	ppgRed	ppgIR	Drowsiness
Heart_Rate	1	0.25	0.48	0.51	-0.5
ppgGreen	0.25	1	0.56	0.41	-0.043
ppgRed	0.48	0.56	1	0.84	-0.62
ppgIR	0.51	0.41	0.84	1	-0.6
Drowsiness	-0.5	-0.043	-0.62	-0.6	1

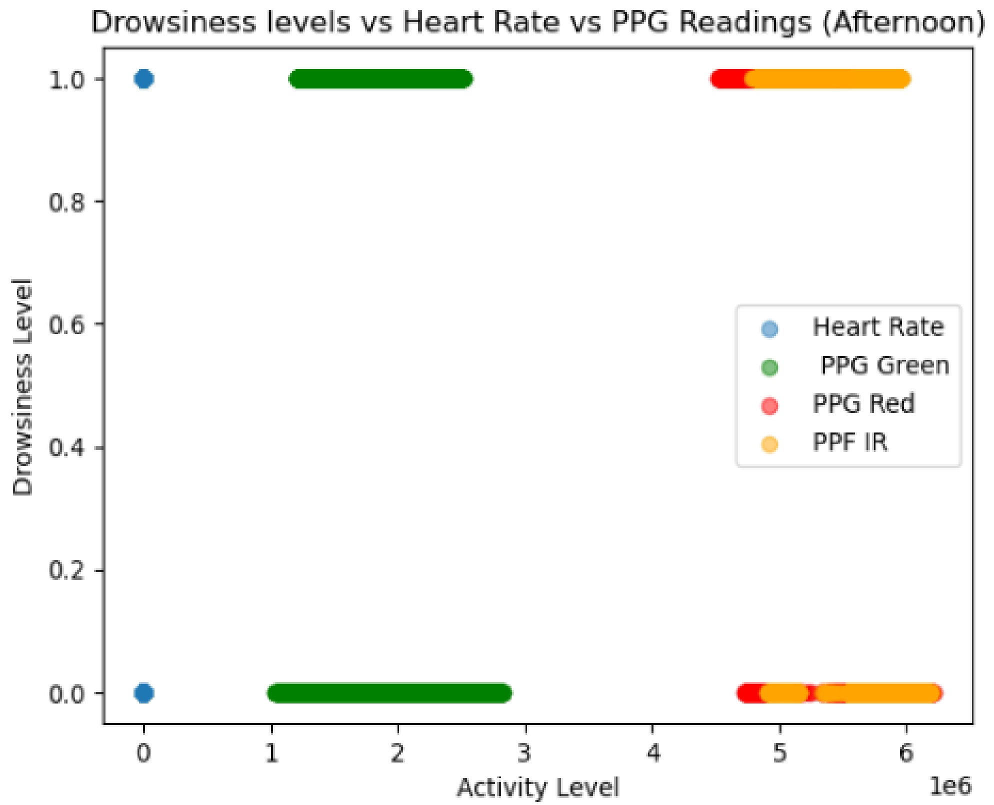


Scatter Plot of ppgGreen Vs Drowsiness

Correlation between Drowsiness and other variables in the Afternoon



```
Correation between drowsiness and Heart rate (Afternoon):0.2226641596779855б
Correlation between drowsiness and ppgGreen(Afternoon):-0.0332603571101783
Correlation between drowsiness and ppgRed(Afternoon):-0.8123097258205565
Correlation between drowsiness and ppgIR (Afternoon):-0.746030609345268
```



- In the afternoon ppg Red and ppgIR levels are inversely proportional to the drowsiness when the you are potentially inactive in the afternoon ppgIR and ppgRed levels go up.
- ppgGreen is having no relation to the drowsiness levels
- In contrast, as the heart rate increases in the afternoon ,the drowsiness increases.

Insights

- There is a very strong positive correlation between ppgRed and ppgIR. They are very much dependent on each other, as the ppgRed value increases the ppgIR value also increases and vice versa.
- Drowsiness is having a negative correlation with Heart Rate, ppgRed and ppgIR and no correlation with ppgGreen.
- As the Heart Rate increases the Drowsiness value decreases which means when the heart Rate is high, the individual is very alert.
- ppgGreen is having a positive correlation with Heart Rate, ppgRed and ppgIR but no correlation with Drowsiness.
- ppgIR is having a very strong correlation with ppgRed, positive correlation with ppgGreen nad Heart Rate but negative correlation with Drowsiness.
- For an alert individual the Heart Rate, ppgRed and the ppgIR value will be high.Not much dependent on ppgGreen value.
- Heart Rate is having a positive correlation with the ppgRed, ppgGreen and ppgRed variables but a negative correlation with Drowsiness.

Recommendations

- It would have been better if more variable like age ,health conditions etc could be considered for the analysis.
- From the dataset we can see that higher the heart rate, more alert the individual is.
- In this dataset the morning ,afternoon were randomly segmented from the dataset . It would have been more insightful if the details were variable at the time of data collection.